Inferences of Species Phylogeny in Relation to Segregation of Ancient Polymorphisms

Chung-I Wu

Department of Biology, University of Rochester, Rochester, New York 14627

Manuscript received May 30, 1990
Accepted for publication October 18, 1990

ABSTRACT

Standard formulas of gene frequency change under genetic drift are used to derive the probability of obtaining incorrect phylogenetic information for three species due to segregation of ancient polymorphisms. This probability depends upon the level of polymorphisms at the time of speciation and is generally quite high unless the two speciation events are far apart in time. If phylogenetic data from multiple loci are available, a likelihood ratio test can be used to reject the null hypothesis in favor of the best phylogeny. The appropriate null hypothesis is either a trichotomy or an alternative phylogeny, depending on the data set. The likelihood ratios required for accepting the best phylogeny are given. These ratios are obtained by exact enumeration when the number of loci is small ($n < 15$) and by an asymptotic approach for larger n's. In general, more than five loci are needed to resolve the species phylogeny.

MOST molecular phylogenetic studies address the question of gene genealogy. In certain cases, the interest has been on the phylogeny of closely related species. One example is the relationships among human, chimpanzee and gorilla (KOOP *et al.* 1986; MAEDA *et al.* 1988; MIYAMOTO *et al.* 1988); another is among *Drosophila simulans, D. mauritiana* and *D. sechellia* (COYNE and KREITMAN 1986). Even if the phylogeny of genes is unambiguous, as that of the β-globin region from human, chimpanzee and gorilla seems to be (see KOOP *et al.* 1989 for the composite data), the phylogeny of the species may still be quite different, as depicted in Figure 1. Imagine that allele A_i is a DNA deletion that is nonrecurrent and irreversible. It is nevertheless quite possible that the gene from species 3 shares this unique character with either that of species 1 or species 2, although the latter two species are more closely related. The cause of the discrepancy is the segregation of (ancient) polymorphisms between nodes 1 and 2.

The first part of this paper addresses the probability that a single character would yield incorrect phylogenetic information. The problem has been addressed by TAJIMA (1983), HUDSON (1983), NEI (1986), PAMILO and NEI (1988) and TAKAHATA (1989), all of whom use the theory of gene coalescence. In this paper, the diffusion theory is used in order to take gene frequency into consideration. In the second part, a likelihood ratio test is proposed for inferring the species phylogeny, if gene genealogies from multiple loci are available. The results are compared with those of previous studies such as FELSENSTEIN (1985).

SINGLE-LOCUS MODEL

We shall first consider a single locus with two or more alleles. We may envisage these alleles as a defined stretch of DNA sequence with certain mutational characteristics, such as insertions/deletions or nucleotide changes. The phylogenetic relationship of the three species is often inferred by sampling one gene from each species, as shown in Figure 1. The question is: When the genes sampled from two of the three species are the same allele, does the gene phylogeny accurately represent the phylogeny of the species? In other words, we have to assess the relative probabilities of pattern α, β_1 or β_2 in Figure 1. We exclude from consideration events that are uninformative about the relatedness of the genes; for example, when all three genes are identical. Depending on the model, such event can still be statistically informative in phylogenetic reconstruction.

Let the frequency of this shared allele, A_i, be p at node 1, which becomes x at node 2 due to drift. Let the time between the two nodes be t generations. We shall measure time in T which has a unit of $2N_e$ generations where N_e is the effective population size, *i.e.*, $T = t/(2N_e)$. The probability $B(p)$ of obtaining incongruent phylogenetic information β_1 or β_2 of Figure 1 is

$$B(p) = p \int_0^1 \phi(p,x;\,T)x(1-x)dx = p^2(1-p)e^{-T}, \quad (1)$$

where $\phi(p,x;\,T)$dx is the probability of transition from p to $(x, x + dx)$ at time T (CROW and KIMURA 1970;

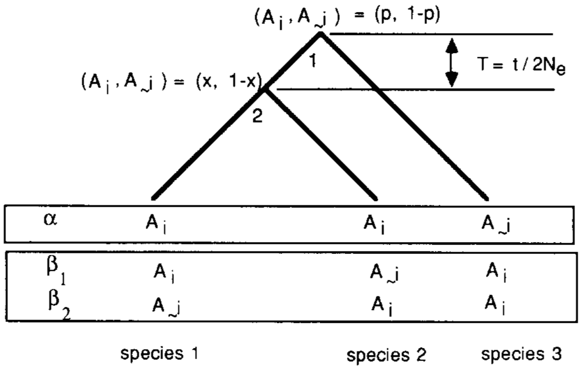

FIGURE 1.—Phylogeny of three species, A_i denotes a particular allele while $A_{\sim i}$ denotes any allele other than A_i. The relative frequency, A_i:$A_{\sim i}$, is p:1-p at node 1 and x:1-x at node 2. One gene from each of the three extant species is sampled. Only the cases where the same allele is sampled twice among the three species are considered in this model.

TABLE 1

Probabilities of phylogenetic incongruence for different allele frequencies (p) and internode times (T)

p	$A(p)$	$B(p)$	$2C(p)$	$2Q(T)$
$T = 0.5$				
0.10	0.0409	0.0055	0.2108	
0.30	0.1208	0.0382	0.3874	
0.67	0.1773	0.0899	0.5034	
0.90	0.0845	0.0491	0.5375	
Eq. (6)				0.4471
NEI's				0.4043
$T = 1.5$				
0.10	0.0719	0.0020	0.0529	
0.30	0.1772	0.0141	0.1369	
0.67	0.2057	0.0331	0.2432	
0.90	0.0880	0.0181	0.2912	
Eq. (6)				0.1515
NEI's				0.1488
$T = 2.5$				
0.10	0.0834	0.0007	0.0174	
0.30	0.1979	0.0052	0.0497	
0.66	0.2161	0.0122	0.1011	
0.90	0.0893	0.0066	0.1297	
Eq. (6)				0.0448
NEI's				0.0547
$T = 3.5$				
0.10	0.0876	0.0003	0.0062	
0.30	0.2056	0.0019	0.0182	
0.67	0.2200	0.0045	0.0391	
0.90	0.0897	0.0024	0.0517	
Eq. (6)				0.0132
NEI's				0.0201

$B(p)$ and $A(p)$ are given in Equations 1 and 2. $C(p) = B(p)/[A(p) + 2B(p)]$. $Q(T)$ is either $1/3\ e^{-T}$ (NEI 1986) or is given in (6).

pp. 383 and 386). The probability $A(p)$ of obtaining congruent phylogenetic information α is

$$A(p) = (1 - p)E(X^2)$$
$$= p(1 - p) - p(1 - p)^2e^{-T} \qquad (2)$$

where $E(X^2)$ is the second moment of gene frequency, x (CROW and KIMURA 1970; p. 336). In Table 1, the values of $A(p)$, $B(p)$ and $2C(p)$ are given where $2C(p)$ [$= 2B(p)/(A(p) + 2B(p))$] is the probability that a locus will give information incongruent with the phylogeny. The probability is conditional on the locus being phylogenetically informative (*i.e.*, two out of the three genes are the same allele). We can see that, as T increases, $2C(p)$ decreases. Naturally, when T approaches 0, $2C(p)$ approaches $\frac{2}{3}$ for all p's. What is interesting is that, given the same T, $2C(p)$ increases with p. For example, $2C(p)$ will drop to about 5% at $T = 1.5$ for $p = 0.1$ whereas, for $p = 0.9$, it will only reach the same level at $T = 3.5$. For the resolution of the human-chimpanzee-gorilla trichotomy, if we assume $N_e = 50{,}000$ between nodes 1 and 2 and a generation time of 10 yr, $T = 1.5$ and 3.5 represent 1.5 and 3.5 million years, respectively.

According to this model, the probability of phylogenetical incongruence is conditional on the value of p at node 1. The distribution of p, $F(p)$, can be assumed to be proportional to the stationary frequency spectrum of the infinite-allele model (KIMURA and CROW 1964);

$$F(p) = M(1 - p)^{M-1}p^{-1}/n_a$$

where $M = 4N_eu$ and $n_a = \int_{1/2N}^{1} M(1 - p)^{M-1}p^{-1}dp$ is the average number of alleles (CROW and KIMURA 1970, p. 455).

The probability of obtaining phylogenetically congruent information (pattern α) for a randomly chosen locus is

Prob(α)

$$= \int_{1/2N}^{1} A(p)F(p)dp$$

$$= \{M(1 - e^{-T}) \int_{1/2N}^{1} (1 - p)^{M+1}dp$$

$$+ M \int_{1/2N}^{1} p(1 - p)^{M}dp\}/n_a$$

$$= M[(M + 2) - (M + 1)e^{-T}]/[(M + 2)(M + 1)n_a].$$

Similarly, the probability of obtaining phylogenetically-incongruent information is

$$\text{Prob}(\beta_1) = \text{Prob}(\beta_2) = \int_{1/2N}^{1} B(p)F(p)dp$$

$$= M\ e^{-T}/[(M + 2)(M + 1)n_a].$$

Therefore, the probability, P, of phylogenetic congruence among all phylogenetically informative loci is

$$P(T) = \text{Prob}(\alpha)/\{\text{Prob}(\alpha)+\text{Prob}(\beta_1)+\text{Prob}(\beta_2)\} \qquad (3)$$
$$= \frac{(M + 2) - (M + 1)e^{-T}}{(M + 2) - (M - 1)e^{-T}}.$$

Similarly,

$$Q(T) = e^{-T}/[(M + 2) - (M - 1)e^{-T}] \qquad (4)$$

is the probability of each phylogenetic incongruence (β_1 or β_2 of Figure 1). In the case of $M = 1$, $P(T) = 1 - \frac{2}{3}e^{-T}$ and $Q(T) = \frac{1}{3}e^{-T}$. The results are identical with the formulas of NEI (1986). In his formulation, $\frac{2}{3}e^{-T}$ is the probability that the genes from species 1 and 2 are less closely related to each other than one of them is to the gene from species 3, regardless of their allelic status.

Mutation between nodes: In the above formulation, mutations between nodes 1 and 2 are not considered. Such mutations would always give rise to phylogenetic congruence. It is thus desirable to account for these internodal mutations by defining a new event, γ. Event γ is like event α of Figure 1, except that $p = 0$ at node 1. The probability of γ is identical with the second moment of gene frequency under irreversible mutation while p approaches 0 (CROW and KIMURA 1970, p. 392) and is given by

$$\text{Prob}(\gamma) = 1 - 2(M + 1)/(M + 2)e^{-M/2T} + M/(M + 2)e^{-(M+1)T}$$

A correction term of Prob(γ) is now introduced into Equation 3 and the new P(T) is given by $\{\text{Prob}(\alpha) + \text{Prob}(\gamma)\}/\{\text{Prob}(\alpha) + \text{Prob}(\beta_1) + \text{Prob}(\beta_2) + \text{Prob}(\gamma)\}$.

Of special interest are the cases when $M \ll 1$. This applies when the allelic status of DNA sequences is defined by a very rare and irreversible mutation, such as a deletion or insertion at a particular nucleotide position. Such nonrecurrent mutational events actually agree best with the assumptions of the infinite-allele model. In addition, the genealogy of these alleles is unambiguous, unlike alleles defined by multiple recurrent events (see LI 1989).

With the correction term and the assumption of $M \ll 1$, we obtain

$$P(T) \approx (1 + T)/(1 + T + 2e^{-T}) \qquad (5)$$

$$Q(T) \approx e^{-T}/(1 + T + 2e^{-T}). \qquad (6)$$

Some representative values of $2Q(T)$ based on Equation 6 as well as NEI's (1986) formula, $\frac{1}{3}e^{-T}$, are given in Table 1. These values form an interesting comparison with the case of $p = \frac{2}{3}$, when a polymorphic locus is most likely to yield phylogenetically incongruent information. (This can be shown by solving for $\partial B(p)/\partial p = 0$ to obtain $p = \frac{2}{3}$ as a local maximum.) When $T = 0.5$, $p = \frac{2}{3}$ will lead to an incorrect inference of the species phylogeny more than 50% of the time [$2C(p)$]. If all values of p are considered, the probability of phylogenetic incongruence [$2Q(T)$'s] is still greater than 40%. Despite very different derivations and interpretations, the values of $2Q(T)$ based on Equation 6 are generally close to $2Q(T) = \frac{2}{3}e^{-T}$, obtained by NEI (1986). In summary, the probability of making an incorrect inference of species phylogeny is unacceptably high unless T is large. If a locus is to yield correct phylogenetic information at the 5% level of significance, the time between node 1 and node 2 has to be about $T = 2.4$ when $M \ll 1$ (Table 1). For the human-chimpanzee-gorilla trichotomy, this means 2.4 million years in the numerical example given previously.

MULTIPLE-LOCUS TEST

It is apparent that one locus alone is insufficient for the resolution of species phylogeny. We may now consider n loci that are not tightly linked (*e.g.*, $N_ec \gg 1$ where c is the recombination fraction between any two loci). We want to accept one of the three phylogenies with the necessary statistical confidence:

Phylogeny A – [(spp 1, spp 2), spp 3]
Phylogeny B – [(spp 1, spp 3), spp 2]
Phylogeny C – [(spp 2, spp 3), spp 1].

Let the number of loci supporting phylogeny A, B or C be a, b or c, respectively, where $a + b + c = n$.

The probability of obtaining the data set (a, b, c) under phylogeny A is

$$G_A(T) = [n!/a!b!c!]P(T)^aQ(T)^{b+c} \qquad (7)$$

where $P(T)$ and $Q(T)$ may be given by equations such as (5) and (6). The probability under phylogeny B and C can be obtained by exchanging a and b and by exchanging a and c of the above formulae, respectively. For each phylogeny, we choose the maximum likelihood estimator (MLE) of T that maximizes $G_A(T)$. This is necessary because each phylogeny is a composite hypothesis with an unspecified internodal length. (The length of other branches only indirectly affects this model, see DISCUSSION.) Because $P(\mathbf{T_A})$ and Q ($\mathbf{T_A}$) are uniquely defined by $\mathbf{T_A}$, the procedure is equivalent to finding the MLE of $P(T)$ and $Q(T)$, provided that $\mathbf{T_A}$ is nonnegative. We therefore solve for $\partial \ln G_A(T)/\partial P|_P = 0$, and obtain

$$P(\mathbf{T_A}) = a/n,$$

$$Q(\mathbf{T_A}) = (1 - a/n)/2 = (b + c)/2n. \qquad (8)$$

The value of $\mathbf{T_A}$ can be obtained from either Equation 3 or 5. Table 2 gives some representative values based on (5). With either equation, $\mathbf{T_A}$ is nonnegative only when $a \geqslant n/3$. Therefore, for $a < n/3$, $\mathbf{T_A} = 0$ and

$$P(\mathbf{T_A}) = Q(\mathbf{T_A}) = \tfrac{1}{3}. \qquad (8')$$

In other words, if the number of loci congruent with a certain phylogeny is less than $\frac{1}{3}$ of the total, the most likely arrangement for that phylogeny is one with a single point of trifurcation.

We may then use the likelihood ratio test to discriminate among the three phylogenies. For convenience, let $a \geqslant b \geqslant c$. We want to reject the second best phylogeny, B, in favor of the best phylogeny, A, if the likelihood ratio, $R_{AB} = G_A(\mathbf{T_A})/G_B(\mathbf{T_B})$, is greater than a threshold value K. Given $a \geqslant b \geqslant c$, only data sets

TABLE 2

Likelihood ratios and levels of significance for $n = 10$ loci (only $a \geq b \geq c$ given)

a	b	c	3 × Prob[a]	R_{AO}	R_{AB}	T_A[b]
10	0	0	$3 \times (1/3)^{10}$	3^{10}		
9	1	0	0.11%	1140		1.85
8	2	0		99.1		1.26
8	1	1	1.02%	99.1		1.26
7	3	0		16.4		0.899
7	2	1	5.91%	16.4		0.899
6[c]	4	0		4.41	4.00	0.618
6	3	1		4.41		0.618
6	2	2	23.0%	4.41		0.618
5[c]	5	0		1.80	1.00	0.375
5[c]	4	1		1.80	1.64	0.375
5	3	2		1.80		0.375
4[c]	4	2		1.10	1.00	0.149
4	3	3		1.10		0.149

[a] This is the cumulative probability for points whose likelihood ratios are greater than R_{AO}, given in the next column. The factor of three is for the sum of the three corners (see Figure 2).

[b] T_A is the MLE of T and is the solution for $P(T) = (1 + T)/(1 + T + 2e^{-T}) = a/n$ (from Equations 5 and 8).

[c] Indicates cases with $b > n/3$ where $R_{AO} > R_{AB}$.

with $b < n/3$ need to be considered in the statistical test if $n \leqslant 15$ (discussed later). This, in turn, means that the second best phylogeny is always phylogeny B with a zero internodal distance for $n \leqslant 15$ (see Equation 8). Thus, discrimination among the three phylogenies is conveniently reduced to testing the null hypothesis of a trifurcation tree against phylogeny A with an internodal length of $\mathbf{T_A}$. Formally, the null hypothesis is

$$H_0\text{: Phylogeny } O\text{-[spp 1, spp 2, spp 3]}$$

against the alternative phylogeny A, when $c \leqslant b \leqslant n/3$. We shall discuss cases with $b > n/3$ later.

It is straightforward to calculate R_{AO} $[G_A(\mathbf{T_A})$ from Equations 7 and 8 and $G_O(0) = n!/a!b!c!\,(\frac{1}{3})^n]$. What we need to determine is the threshold value, K, above which phylogeny O will be rejected in favor of phylogeny A. As there are three ways to reject O, we accept A if the probability of $R_{AO} > K$ is less than $0.05/3 \approx 0.017$ under the null hypothesis. The statistic, $2\ \ln(R_{AO})$, has an asymptotic chi-square distribution with one degree of freedom (HOEL, PORT and STONE 1972). The asymptotic value for K is, therefore, $K_{\text{asym}} = e^{5.7/2} \approx 17.3$, where $5.7 = \chi^2_{0.017}$ with one degree of freedom. When n is sufficiently large, we reject the null hypothesis in favor of phylogeny A at the 5% significance level if

$$R_{AO} = (3/n)^n\{a^a[(b + c)/2]^{b+c}\} \geqslant K_{\text{asym}} = 17.3. \quad (9)$$

In practice, we have to know how quickly the asymptotic chi-square distribution is approached as n gets larger. More importantly, because we rarely have molecular data of more than 5 loci from all three species, it is necessary to calculate the exact K values for small n's.

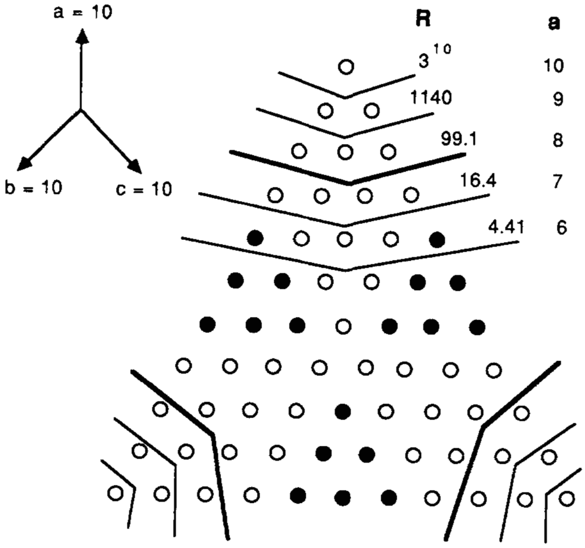

FIGURE 2.—The total sample space for $n = 10$ loci. Each circle represents a vector (a, b, c) where a is the number of loci supporting phylogeny A, b supporting B and c supporting C with $a + b + c = n$. The apex at the top represents (10, 0, 0) and the bottom (0, b, c), while the apex of the left corner represents (0, 10, 0) and of the right corner, (0, 0, 10), as indicated by the arrows. R is the likelihood ratio of Equations 9 for all the points between the appropriate contour lines. Points above the thick lines are the region for rejecting the null hypothesis in favor of the best phylogeny. The filled circles represent samples where the number of loci supporting the second best phylogeny is greater than $n/3$. If filled circles fall in the region of rejection, acceptance of the best phylogeny requires a different criterion (R_{AB}), specified in Equation 9′

The exact enumeration of the distribution of R_{AO} was carried out for n up to 40. It was done by summing up $\text{Prob}(a,b,c) = n!/a!b!c!\ (\frac{1}{3})^n$ according to the descending values of $R_{AO}(a,b,c)$'s. An example for $n = 10$ is given in Fig. 2 where all the data points are given (see also FELSENSTEIN 1985). Contour lines for different values of R are also given. The cumulative probabilities for data points above a contour line (*i.e.*, toward the corner), including points in all *three* corners of Figure 2, are given in Table 2. For $n = 10$, the probability of $R > 16$ under the null phylogeny is 5.9%, slightly larger than 5%. Therefore, a phylogeny is significantly better than the null phylogeny only when there are eight or more out of 10 loci supporting it. Data points that will result in the acceptance of the alternative phylogeny are those above the thick contour lines in Figure 2. The MLE of the internodal distance under the best phylogeny is also given as $\mathbf{T_A}$ in Table 2. For $n = 10$, $\mathbf{T_A} = 0.899$ is not significantly different from 0.

Table 3 summarizes the results of the exact enumeration of R_{AO} for n up to 40. To reject the null hypothesis in favor of a particular phylogeny, we need

TABLE 3

Threshold values of likelihood ratios, *K*, for accepting the best phylogeny over trichotomy

n	L^a	K	$3 \times \text{Prob}^b$	K_{asym}	K/K_{asym}
4	4	3^4	$3 \times (1/3)^4$		
5	5	3^5	$3 \times (1/3)^5$		
6	6	3^6	$3 \times (1/3)^6$		
7	6	61	2.06%	34.8	1.78
8	7	161	0.78%	127	1.26
9	7	41	2.48%	32.2	1.29
10	8	99	1.02%	72.9	1.35
11	8	35	2.65%	30.7	1.14
12	9	77	1.16%	64.7	1.20
13	9	32	2.65%	30.7	1.06
14	10	68	1.21%	62.8	1.09
15	10	32	2.55%	31.5	1.01
>15		17.2	5%	17.2	

For $n > 15$, K at the 5% level is the asymptotic value, 17.2.

[a] L is the minimal number of loci that support phylogeny A, required to give a $R_{AO} \geq K$. For $n > 15$, the numbers for n and L are: 16(10), 17–18 (11), 19–20 (12), 21–22 (13), 23–25 (14), 26–27 (15), 28–29 (16), 30–32 (17), 33–34 (18), 35–36 (19), 37–39 (20), 40 (21).

[b] $3 \times \text{Prob}\ (R_{AO} \geq K)$.

to have data showing a minimal number of loci (L) congruent with that phylogeny. The threshold values, K, and the probabilities of $R_{AO} > K$, summed over all three corners, are also given. In general, the values of L are either identical with or larger by one than those of FELSENSTEIN (1985). The levels of significance are not exactly 5% because of the discrete nature of the data. For example, as shown in Table 2 and Figure 2, the level of significance for $n = 10$ is either 1.02% or 5.91%. The highest available level below 5% is chosen. A simple calculation shows that at least 4 loci are needed to resolve the species phylogeny. For $n = 4$ to 6, a phylogeny is significantly better than others only when every locus supports it. For $n \geqslant 7$, K fluctuates appreciably, due mainly to the different levels of significance. In the column under K_{asym} are shown the K values at the corresponding level of significance under the asymptotic chi-square distribution. The ratios of K/K_{asym} in the last column show the asymptotic behavior of K which approaches the asymptote at $n \geqslant 15$. For $n = 7$ to 15, rejection of the null hypothesis requires an R_{AO} value larger than the asymptotic K. For larger n's, Equation 9 should be sufficient.

The above discussion applies only if $b \leqslant n/3$. Data sets with $b > n/3$, represented by filled circles in Figure 2, require some special considerations. In these cases, the null phylogeny to be rejected should not be phylogeny O; rather it is phylogeny B with a nonzero internodal length. The likelihood ratios, R_{AB}'s, are smaller than R_{AO}'s as shown in Table 2 for the data sets with $b > n/3$. This is because the probability of obtaining the data set is higher under phylogeny B than under O. In other words, A is not as significantly better than phylogeny B as it is than O, indicated by the R_{AO} values. This does not present a problem as long as we do not reject O in favor of A. In other words, if phylogeny A is not significantly better than O, it is certainly not better than B. In all cases of $n \leqslant 15$, $R_{AO} < K$ whenever $b > n/3$ because the minimal number (L in Table 2) for $R_{AO} > K$ is always $a \geqslant 2n/3$ (hence $b < n/3$). For $n \leqslant 15$, we do not have to test phylogeny A against the second best phylogeny, B, even when $b > n/3$.

When n is larger than 15, further enumeration is perhaps necessary. For example, an observation of $(a, b, c) = (21, 18, 1)$ will lead to the acceptance of phylogeny A over phylogeny O but phylogeny A is, in fact, not statistically better than B with $P(\mathbf{T_B}) = 18/40$. A situation with $a \geqslant L$ and $b \geqslant n/3$ is expected to be rare under any phylogeny and the appropriate likelihood ratio can be obtained as

$$R_{AB} = 2^{a-b}[a^a(b+c)^{b+c}]/[b^b(a+c)^{a+c}],\ \text{if}\ b > n/3 \qquad (9')$$

similar to Equation 9. One may argue that, since this situation arises only when $n > 15$ where the asymptotic chi-square distribution is appropriate, the criterion for accepting A over B is $R_{AB} \geqslant K_{\text{asym}} = 17.3$. However, the degree of freedom in those cases is not one. Exact enumeration as done in Fig. 2, but based on the appropriate $P(\mathbf{T_A})$, $Q(\mathbf{T_A})$, $P(\mathbf{T_B})$ and $Q(\mathbf{T_B})$ values, will be necessary to determine the correct K for any given data set with $b > n/3$ and $n > 15$.

DISCUSSION

Segregation of ancient polymorphisms is a serious confounding factor in the inference of species phylogeny. In fact, the data appear to reflect that process. For example, KOOP *et al.* (1986) and MAEDA *et al.* (1988) found two deletions, about 4 kb apart in the β-globin region, that are shared exclusively between human and chimpanzee. Common alleles shared by chimpanzee and gorilla have also been reported for a mitochondrial DNA deletion (HIXON and BROWN 1986) and the involucrin gene (DJIAN and GREEN 1989). The mutual sharing of alleles is no less difficult a problem in inferring the phylogeny of the three sibling species in the *D. melanogaster* group (COYNE and KREITMAN 1986).

The model presented here assumes that the genealogy of genes is unambiguous. The message of this study is a simple one: Even if the genealogy of genes is unambiguous, it still takes a large number of congruent loci to resolve the phylogeny of species. A straightforward way of inferring gene genealogy is to rely on "unique and derived" characters, such as the shared deletions reported by KOOP *et al.* (1986) and MAEDA *et al.* (1988) in the β-globin region of human and chimpanzee. Another example may be the insertions of elements like the *Alu* repetitive sequences

(Hwu *et al.* 1986). Since the genealogy is not always unambiguous (perhaps for want of such uniquely derived characters), the task of correctly inferring species phylogeny is even more difficult than suggested here. For instance, although the genealogy of the β-globin genes studied by Koop *et al.* (1986), Maeda *et al.* (1988) and Koop *et al.* (1989) is convincing based on the two shared deletions, Li (1989) did not find their genealogy statistically convincing when only nucleotide substitutions were studied.

It is important to note that the model deals with genes that segregate as independent units. Linked multiple changes that distinguish two segregating alleles are considered one character. Exactly how much recombination between two loci is necessary for them to be considered independent is beyond the scope of this study. The two deletions in the β-globin region discussed above are perhaps best treated as part of the same segregating unit, especially in light of the finding of strong linkage disequilibrium over a long stretch of DNA in humans (*e.g.*, Murray *et al.* 1984).

Strictly speaking, the model depends only on the length between nodes 1 and 2. Of course, very long branches leading to the extant species will accumulate additional changes that would obscure the phylogenetic information at node 2. But the measurement of T in $2N_e$ generations only assumes that, between the two nodes, constant population size is maintained. The branch lengths leading from node 2 to species 1 and 2 and from node 1 to species 3 are important if multiple alleles are sampled from each species (Takahata 1989). This is why sampling many genes (alleles) of the same locus will not help resolve the human-chimpanzee-gorilla trichotomy. Cann *et al.* (1987) have shown that the coalescence time of human mitochondrial DNA is around 200,000 yr. The coalescence time for nuclear genes is perhaps about 400,000 yr. This is much below the divergence time of the great apes, which is not likely to be less than 5 million years. In other words, multiple alleles drawn from the extant ape species would actually represent a sample of only one single allele at the time of speciation. The situation is perhaps quite different among *D. simulans*, *D. mauritiana* and *D. sechellia*. Coyne and Kreitman (1986) found some shared polymorphisms between these species. This suggests that a multiallele sample may yield additional information on the species phylogeny (Pamilo and Nei 1988; Takahata 1989).

This single-locus model differs from previous studies such as Pamilo and Nei (1988) in that it brings the allele frequency into consideration. In the analyses of empirical data, variations in the allelic status within and between species are critical but have not been incorporated into previous studies. For example, genes of a conservative locus, such as 5S RNA, may have a genealogy as predicted by the coalescent time theory but, in the absence of any variations, such a genealogy will be completely transparent to an observer. It is for this reason that one might expect the allele frequency, or a parameter governing the allele frequency (such as $4N_eu$), to be incorporated in the analysis.

There are indeed three processes involved: the separation of species, the genealogy of genes and the introduction of detectable variations into these genes. The first two processes can be decoupled from the empirical practices as long as certain assumptions are true (such as the neutrality and a constant population size). The last one depends very much on the empirical practices, such as how one determines different alleles, and is critical to an observer for the inference of phylogeny. In this model, the third process is considered under certain assumptions. Equations 3 and 4 assume that mutations occurring between the two nodes do not add more power to the inference of gene genealogy. This may be true if the samples from species 1 and 2 (as in Figure 1) are A_i, which acquired an additional mutation between the two nodes, while the sample from species 3 is A_j. In that case, the allelic differences at node 1 are sufficient for the correct genealogical inference. This assumption is not valid if $M \leq 1$ when the allelic difference is either completely present or completely absent at node 1. Therefore, Equations 5 and 6 were derived to incorporate internodal mutations.

The likelihood ratio for multiple-locus data provides a well defined hypothesis test. Here the appropriate null hypothesis is the second best phylogeny, which is tested against the best phylogeny as the alternative hypothesis. To summarize (let $a \geq b \geq c$): When $n \leq 15$, the null hypothesis is the trichotomy and the best phylogeny will be accepted if $R_{AO} = (3/n)^n \{a^a[(b+c)/2]^{b+c}\} \geq K$ (given in Table 3). When $n > 15$ and $b < n/3$, the test is as in Equation 9 using the asymptotoic $K = 17.3$. When $n > 15$ and $b > n/3$, the null hypothesis is phylogeny B (its internodal length given by Equation 8) and the likelihood ratio is given in (9′). The results are in general agreement with Felsenstein's (1985) study although the tests are quite different.

Finally, since it is possible to estimate $T = t/(2N_e)$ given good cladistic data, it may be possible to estimate N_e if an independent estimate of t [*e.g.*, that by Sibley and Ahlquist (1984)] is used.

I thank Joe Felsenstein and Dick Hudson for many helpful suggestions. I also thank James F. Crow, Andy Clark, Jack Werren, Norman Johnson and Mike Palopoli for helpful comments. This study is supported by a National Institutes of Health grant (GM39902) and a National Down Syndrome Society Fellowship.

LITERATURE CITED

CANN, R. L., M. STONEKING and A. C. WILSON, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

COYNE, J. A., and M. KREITMAN, 1986 Evolutionary genetics of two sibling species, *Drosophila simulans* and *D. sechellia*. Evolution **40:** 673–691.

CROW, J. F., and M. KIMURA, 1970 *An Introduction to Population Genetics Theory.* Burgess Publishing Co., Minneapolis, Minn.

DJIAN, P., and H. GREEN, 1989 Vectorial expansion of the involucrin gene and the relatedness of the hominoids. Proc. Natl. Acad. Sci. USA **86:** 8447–8451.

EWENS, W. J., 1979 *Mathematical Population Genetics.* Springer-Verlag, Berlin.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies with a molecular clock. Syst. Zool. **34:** 152–161.

HIXON, J. E., and W. M. BROWN, 1986 A comparison of the small ribosomal RNA genes from the mitochondrial DNA of the great apes and human: sequence, structure, evolution and phylogenetic implications. Mol. Biol. Evol. **3:** 1–18.

HOEL, P. G., S. C. PORT and C. J. STONE, 1972 *Introduction to Statistical Theory.* Houghton Mifflin, Boston.

HUDSON, R. R., 1983 Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HWU, H. R., J. W. ROBERTS, E. H. DAVIDSON and R. J. BRITTEN, 1986 Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. Proc. Natl. Acad. Sci. USA **83:** 3875–3879.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KOOP, B. F., M. GOODMAN, P. XU, K. CHAN and J. L. SLIGHTOM, 1986 Primate eta-globin DNA sequences and man's place among the great apes. Nature **319:** 234–238.

KOOP, B. F., D. A. TAGLE, M. GOODMAN and J. L. SLIGHTOM, 1989 A molecular view of primate phylogeny and important systematic and evolutionary questions. Mol. Biol. Evol. **6:** 580–613.

LI, W. -H., 1989 A statistical test of phylogenies estimated from sequence data. Mol. Biol. Evol. **6:** 424–435.

MAEDA, N., C. I. WU, J. BLISKA and J. RENEKE, 1988 Molecular evolution of higher primates: intergenic structure, rate and pattern of DNA changes and origin of simple sequences. Mol. Biol. Evol. **5:** 1–20.

MIYAMOTO, M. M., B. F. KOOP, J. L. SLIGHTOM, M. GOODMAN and M. R. TENNANT, 1988 Molecular systematics of higher primates: genealogical relationships and classification. Proc. Natl. Acad. Sci. USA **85:** 7627–7631.

MURRAY, J. C., K. A. MILLS, C. M. DEMOPULOS, S. HORNUNG and A. G. MOTULSKY, 1984 Linkage disequilibrium and evolutionary relationship of DNA variants at the serum albumin locus. Proc. Natl. Acad. Sci. USA **81:** 3486–3490.

NEI, M., 1986 Stochastic errors in DNA evolution and molecular phylogeny in *Evolutionary Perspectives and the New Genetics*, Alan R. Liss, New York.

PAMILO, P., and M. NEI, 1988 Relationships between gene trees and species trees. Mol. Biol. Evol. **5:** 568–583.

SIBLEY, C. G., and J. E. AHLQUIST, 1984 The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J. Mol. Evol. **20:** 2–16.

TAJIMA, F., 1983 Evolutionary relationships of DNA sequences in finite populations. Genetics **105:** 437–460.

TAKAHATA, N., 1989 Gene geneaology in three related populations: consistency probability between gene and population trees. Genetics **122:** 957–966.

Communicating editor: D. CHARLESWORTH